# Machine Learning Algorithms to Predict Students' Programming Performance: A comparative Study

Sivasakthi M[1] and Pandiyan M[2]

[1]*Assistant Professor*
*Dept. of Comp. Science and Application*
*CSH, SRM IST, Vadapalani*
*Chennai-600 026, Tamilnadu, India*

[2]*Assistant Professor*
*Dept. of Comp. Science*
*CSH, SRM IST, Kattankulathur*
*Chennai-603 203, Tamilnadu, India*

[1]*sivasakm1@srmist.edu.in*, [2]*pandiyam@srmist.edu.in*

***Abstract.*** *Research on students' academic performance is swiftly greater than ever in the field of education, especially students' performance in programming is crucial. Predicting the performance of students in programming using machine learning algorithms and comparing them to suggest a best model will bestow benefits to the students and teachers. Thus a study has been carried out to suggest a best model for students' learning in program by comparing the experiments results of Naïve Bayes and Decision Tree, K-Nearest Neighbor, Support Vector Machine and Random Forest algorithms. Data collection, pre-process and classification process are the sequence of steps for building and comparing the models. Test results indicate that Naïve Bayes confers the best accuracy of 91.02% and SVM algorithm has a high accuracy of 88.77%.*

**Keywords: Machine learning algorithms; Classification; Prediction; Computer rogramming**

## 1. INTRODUCTION

Programming is a crucial part of computer courses. Improving the quality of writing programming code is always demand in IT industry. There is a gap between students knowledge in programming and industry expectations. The students' programming performance prediction is an incredibly key concern for improving their programming knowledge. Even though many studies have investigated for prediction of programming performance, if different parameters with other educational settings then, it can be difficult to apply, considerably, how the programming languages are being taught and the evaluation formation used. The students' programming learning and excellence of computer courses can be improved by predicting students programming performance early [1]. Machine learning algorithms facilitate to predict the programming performance of students [2].

Machine learning research proposes the important technique for data classification. Classifying the data is depends on grouping of data according to the predicting attributes [3]. Classification represents by variety of models. Some trendy classification algorithms are Naïve Bayesian classifier, K-Nearest Neighbors classifier, Random Forest, Decision Tree Algorithms, and Support Vector Machine. The key objective of this paper is to classify the students' performance in programming and to compare the efficiencies of various machine learning algorithms through which predicting the Students' programming performance.

Plenty of studies have been done on Comparing several machine learning algorithms to get the best predictions. Among those comparing studies; Simple Logistic Classifier and SVM algorithms to predict athlete's win [4], comparative analysis between Support Vector Machine and K-Nearest Neighbor classifier for EMG signal classification [5], compare K-Nearest Neighbor, Support Vector Machine, and Random Forest algorithms for facial expression classification [6].

This paper prioritizes models to predict the effectiveness of student programming using the using five machine learning algorithms; of Naïve Bayes and Decision Tree, K-Nearest Neighbor, Support Vector Machine and Random Forest. The process of obtaining result consists data collection; pre–processing for quality of data and Classification process is for applying Classification algorithms.

## II. RELATED WORKS

Researches on predictive analyses were conducted to identify the students' academic performance by various researchers [7-11], significantly a study has been accomplished in identifying the students' programming performance [12]. Performance of students' programming depends on several input parameters like Gender, medium of Instruction, Higher secondary with computer science, parent's work, Nativity, Programming Aptitude, Problem solving skills, E-Learning usage [13-17]. We bring in 'interest in programming', 'Plan After Graduation' and 'First Graduate' attributes in this study because we strongly believe that the students interest with respect to programming and their plan after graduation certainly will reflects in their programming performance and first graduate unquestionably will replicate in their learning. Table-1 demonstrates the summary of research works related to our study.

Research for predicting student performance had been done earlier. Among them comparing Bayesian algorithm and Decision Tree [18], compare Apriori and K-Means algorithms [19], and compare Neural Network, SVM, and Decision Tree algorithms [20], compare the KNN, SVM, and Decision Tree algorithms[21], compare the Naïve Bayes, Decision Tree, SVM, KNN and Random Forest & deep neural network algorithms in student performance[22]. Nobody compared the Naïve Bayes, Decision Tree, SVM, KNN and Random Forest for predicting programming performance of students. Therefore the idea for doing research arises with aim to compare the above machine learning algorithms to suggest the best model for predicting programming performance of students.

**Table 1. Review of Literature**

| Year | Paper | Algorithms | Best Algorithm |
|---|---|---|---|
| 2011 | Huang et al, [23] | Linear Regression, Artificial Neural Network, Radial Basis NN, NN, Support Vector Machine. | Support Vector Machine |
| 2012 | Livieris et al, [24] | Artificial Neural Network, Decision Tree, Naïve Bayes, Support Vector Machine | Artificial Neural Network, Support Vector Machine |
| 2014 | Arsad et al, [25] | Linear Regression, Artificial Neural Network | Linear Regression, Artificial Neural Network |
| 2014 | Gray et al, [26] | DT, NB, Logistic Regression, SVM, KNN, ANN | NB, SVM, KNN, |
| 2016 | Hamsa et al,[27] | FGA & Decision Tree | Decision Tree more strict than FGA |
| 2017 | Ihsan A. & Ashraf Y.A. [9] | Naïve Bayesian, KNN | Naïve Bayesian |
| 2019 | Slamet Wiyono & Taufiq Abidin [21] | KNN, SVM, DT | DT |
| 2019 | H.M. Rafi Hasan et al, [8] | KNN, DT,SVC, Random forest, Gradient boost, LDA | DT |
| 2019 | Hussein Altabrawee et al, [10] | ANN, Logistic regression, Naïve Bayes, DT | ANN |
| 2019 | V. Vijayalakshmi & K. Venkatachalapathy[22] | Decision Tree, Naïve Bayes, Random Forest, Support Vector Machine, K-Nearest Neighbor and Deep neural network | Deep neural network |
| 2020 | Vairachilai S et al, [12] | SVM, Naïve Bayes, DT | Naïve bayes |
| 2021 | J. Dhilipan et al, [11] | Binomial Logical regression, Decision Tree, Entropy, KNN | Binomial Logical regression |

## III. IMPLEMENTATION

Five machine learning algorithms have been used to predict students' performances in Python Programming on collected dataset. The Anaconda's Spider IDE has been used for experiments. The data set has applied for pre processing and classification process has done subsequently those results are compared to exemplify the best model. The researcher seek to categorize the students based on their attributes in to whether first class (60% and above) or not.

### A. Data set and Preprocessing

The data have been collected from students who have been enrolled in the affiliated colleges of University of Madras. According to their curriculum in first year and first semester, studying a subject called 'Problem solving using python'.

The data collection was done through questionnaire. In this regard 535 students' responses were received out of which 490(202 male, 288 female) records with 17 attributes have been considered for processing. Table-2 shows the selected attributes for our research work. Cleaning, integration and reduction are the special techniques used to pre-process the data.

**Table 2. Data Set**

| S. No | Attribute | Coding |
|---|---|---|
| 1 | Gender | Binary: 1-Yes, 0-No |
| 2 | Higher Secondary studied Com.Sci | Binary: 1-Yes, 0-No |
| 3 | Medium of Instruction in HSc | Ordinal: from 1 to 6 |
| 4 | Type of Board | Ordinal: from 1 to 3 |
| 5 | Nativity | Nominal: from 1 to 4 |
| 6 | Are You First Graduate | Binary: 1-Yes, 0-No |
| 7 | Is your father working? | Binary: 1-Yes, 0-No |
| 8 | Plan After Graduation | Ordinal: from 1 to 3 |
| 9 | Are you interest in programming? | Binary: 1-Yes, 0-No |
| 10 | Prior Programming Experience | Binary: 1-Yes, 0-No |
| 11 | Mathematics Mark in HSc | Ordinal: from 1 to 4 |
| 12 | Programming Aptitude | Ordinal: from 1 to 5 |
| 13 | Problem solving skills | Ordinal: from 1 to 5 |
| 14 | E-Learning usage | Ordinal: from 1 to 4 |
| 15 | Grade in python at College | Ordinal : from 1 to 6 |
| 16 | Score in python at Test | Binary: 1-First class, 0-Not a First class |

The quality of data preprocessing is made by choosing key attributes. We dropped unnecessary columns such as Email-ID, Register number, Age from the dataset because these three columns will not make any

disparity in the preferred result. Then dataset was pervaded to produce the graph for correlation. On or after the correlation graph we can articulate that test mark is by and large depends on the grade which they obtained in that picky subject and 15 attributes. So our input parameters are such as mentioned above.

### B. Evaluation metrics

To calculate the performances of the different machine learning algorithms various matrices have been used in this study. They are Precision, Recall,F-score and Accuracy. These can be depicted from confusion matrix.

## IV. RESULT AND DISCUSSIONS

In this study, the data set is splitted into two parts with the ratio of 70:30. That is 70% of data for training and 30% of data for testing. Table 3 shows the confusion matrix of Machine Learning Algorithms. After the confusion matrix generated its comparison have done which is shown in table 4.

**Table 3. Confusion Matrix**

| N=490 | | Actual Values | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Naïve Bayes(NB) | | Decision Tree(DT) | | K-Nearest Neighbor(KNN) | | Support Vector Machine(SVM) | | Random Forest (RF) | |
| | | Active | Non-Active | Active | Non-Active | Active | Non-Active | Active | Non-Active | Active | Non-Active |
| Predicted Values | Active | 225 | 19 | 219 | 34 | 212 | 35 | 220 | 25 | 210 | 34 |
| | Non-Active | 25 | 221 | 39 | 198 | 42 | 201 | 30 | 215 | 27 | 219 |

The true prediction on active and on Non-active and false prediction on active and Non-active classification has done on five machine learning algorithms to compare the classification accuracy. It can be seen that the Naïve Bayes can predict correctly 225 active students and 221 non-active students, and Decision Tree can just predict 219 active students and 198 non-active students. K-Nearest Neighbor algorithm predict 212 active students and 201 non-active students.220 active students and 215 non-active students were correctly predicted by Support Vector Machine, while Random Forest predict 210 students as active and 219 students as non-active. Before testing, the comparison of classification accuracy is shown in Fig.1.

**Table 4. Comparison of Confusion Matrix**

| Prediction | | NB | DT | KNN | SVM | RF |
|---|---|---|---|---|---|---|
| Active | Yes | 92% | 87% | 86% | 90% | 86% |
| | No | 8% | 13% | 14% | 10% | 14% |
| Non-Active | Yes | 90% | 84% | 83% | 88% | 89% |
| | No | 10% | 16% | 17% | 12% | 11% |

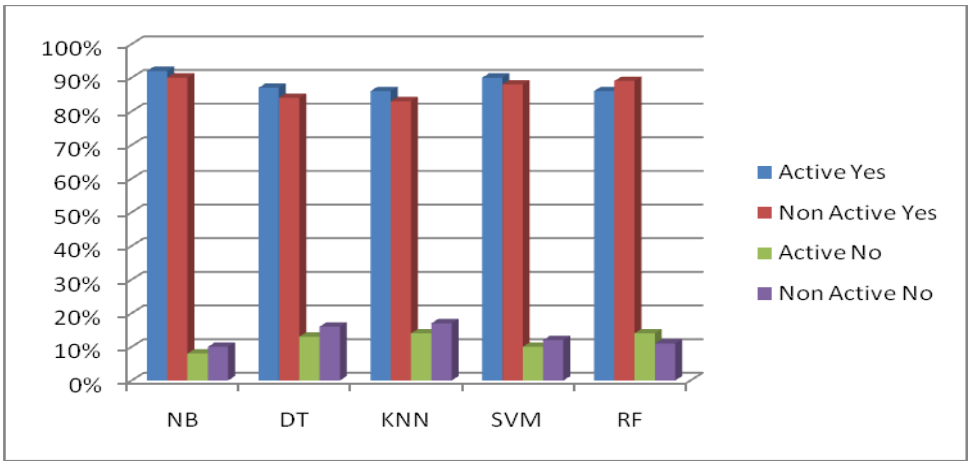


**Fig.1 Comparison of testing accuracy**

In our tests, recall represents proportion of the related information to that class and then its true classification, precision represents the percentage of data that is categorized accordingly. F-score represent the harmonic mean. Percentage of correctly classified instances is accuracy. These are shown in Table 5. The high percentage represents the strong correlation between the features, which influences the student's programming performance.

**Table 5. Results of the Machine Learning Algorithm**

| Machine Learning Algorithms | Size of Dataset | Recall | Precision | F-score | Accuracy |
|---|---|---|---|---|---|
| Naïve Bayes | 490 | 90 | 92.21 | 91.09 | 91.02 |
| Decision Tree | 490 | 84.89 | 86.57 | 85.72 | 85.10 |
| K-Nearest Neighbor | 490 | 83.46 | 85.82 | 84.62 | 84.28 |
| Support Vector Machine | 490 | 88 | 89.79 | 88.88 | 88.77 |
| Random Forest | 490 | 88.60 | 86.06 | 87.31 | 87.55 |

The result shows the NB classification method bestows the improved accuracy (91.02%) when comparing to DT classifier which confers the accuracy of 85.10%, KNN predicts 84.28%, 88.77% by SVM followed by 87.55% by RF. Thus the best algorithm to predict students' programming performance is NB. The next best algorithm is SVM, followed by RF, DT and KNN.

## V. CONCLUSION

Early prediction of students in programming will help them to pick up for enhanced performance. In this study the researcher proposed system for predicting the programming performance of students. The model has been trained and tested using various machine learning algorithms. They are Naïve Bayes and Decision Tree, K-Nearest Neighbor, Support Vector Machine and Random Forest. Lastly compared the results of five algorithms out of which Naïve Bayes provide the best with 91.02% as accuracy. in addition it is confirmed by Susan Bergin's research [28].

**REFERENCES**


1. *V. S. Warke and R. S. Kamath "Data Mining Approach for the Analysis of Performance Based Appraisal System of Selected Teachers in Kolhapur City," no. Iv, pp. 1–6, 2016.*
2. *Sivasakthi M, K. R. Anantha Padmanabhan, "Prediction of Students Programming Performance using Naïve Bayesian and Decision Tree", 2nd International Conference on Soft Computing for Security Applications, April 2022.*
3. *P. Kaur, M. Singh, and G. Singh, "Classification and prediction based data mining algorithms to predict slow learners in education sector," Procedia - Procedia Comput. Sci., vol. 57, pp. 500–508, 2015.*
4 *E. Rainarli and A. Romadhan, "Perbandingan Simple Logistic Classifier dengan Support Vector Machine dalam Memprediksi Kemenangan Atlet," J. Inf. Syst. Eng. Bus. Intell., vol. 3, no. 2, p. 87, 2017.*
5 *Y. Paul, V. Goyal, and R. A. Jaswal, "Comparative analysis between SVM & KNN classifier for EMG signal classification on elementary time domain features," 4th IEEE Int. Conf. Signal Process. Comput. Control. ISPCC 2017, vol. 2017–Janua, pp. 169–175, 2018.*
6 *R. A. Nugraheni and K. Mutijarsa, "Comparative Analysis of Machine Learning KNN, SVM, and Random Forests Algorithm for Facial Expression Classification," in ISEMANTIC, 2016, pp. 163–168.*
7. *F. Haghanikhameneh, P. H. Shariat Panahy, N. Khanahmadliravi, and S. A. Mousavi, "A comparison study between data mining algorithms over classification techniques in Squid dataset," Int. J. Artif. Intell., vol. 9, no. 12 A, pp. 59–66, 2012.*
8. *H.M. Rafi Hasan et al, Machine Learning Algorithm for Student's Performance rediction, IEEE, 10th ICCCNT 2019 July 6-8, IIT – Kanpur, 2019*
9. *Ihsan A. Abu Amra et al, Students performance prediction using KNN and Naïve Bayesian, 2017 8th International Conference on Information Technology (ICIT)*
10. *Hussein Altabrawee et al, Predicting Students' Performance Using Machine Learning Techniques, Journal of University of Babylon, Pure and Applied Sciences, Vol.27, no.1: 2019*
11. *J. Dhilipan et al, Prediction of Students Performance using Machine Learning, IOP Conf. Series: Materials Science and Engineering 1055, 2021*
12. *Vairachilai S et al, Student's Academic Performance Prediction Using Machine Learning Approach, International Journal of Advanced Science and Technology Vol. 29, no. 9s, pp. 6731-6737, 2020*
13. *A. F. El Gamal , An Educational Data Mining Model for Predicting Student Performance in Programming Course, International Journ a l o f Computer Applications, Vol. 70 – no. 17, 2013*
14. *M Sivasakthi, Study of learning difficulties in concurrent programming of OOPs using Java for the students of Computer Science and Engineering, Doctoral thesis, University of Mardras, 2013*
15. *Edin Osmanbegović, Mirza Suljić and Hariz Agić , Determining dominant factor for students performance prediction by using data mining classification algorithms, Tranzicija ,Vol. 16, no. 34, pp. 147-158, 2014*
16. *Sivasakthi M. classification and prediction based data mining algorithms to predict students" introductory programming performance. Proceedings of the International Conference on Inventive Computing and Informatics IEEE Xplore, CFP17L34-ART, Nov. 2017.*
17. *M Sivasakthi, Determining the Central Factors for 'Prediction of Students Programming Performance using Data Mining Algorithms', Vol. 05 – no. 3, PP. 122-128,2018*

18 A. U. Khasanah and Harwati, "A Comparative Study to Predict Student's Performance Using Educational Data Mining Techniques," IOP Conf. Ser. Mater. Sci. Eng., vol. 215, p. 012036, Jun. 2017.

19 G. S. Gowri, R. Thulasiram, and M. A. Baburao, "Educational Data Mining Application for Estimating Students Performance in Weka Environment," IOP Conf. Ser. Mater. Sci. Eng., vol. 263, no. 3, 2017.

20 M. Ciolacu, A. F. Tehrani, R. Beer, and H. Popp, "Education 4. 0 – Fostering Student' s Performance with Machine Learning Methods," in SIITME, 2017.

21. Slamet Wiyono & Taufiq Abidin, Comparative Study of Machine Learning Knn, Svm, And Decision Tree Algorithm To Predict Student's Performance", International Journal of Research-Granthaalayah", vol.7, no 1, 2019.

22. V. Vijayalakshmi & K. Venkatachalapathy, "Comparison of Predicting Student's Performance using Machine Learning Algorithms", I.J. Intelligent Systems and Applications, Vol.12, pp 34-45, 2019

23. S. Huang and N. Fang, "Work in progress: Early prediction of students' academic performance in an introductory engineering course through different mathematical modeling techniques," Proc. - Front. Educ. Conf. FIE, vol. 1, pp. 3–4, 2012.

24. I. E. Livieris, K. Drakopoulou, and P. Pintelas, "Predicting students' performance using artificial neural networks," Proc. 8th Pan-Hellenic Conf. "Information Commun. Technol. Educ., pp. 28–30, 2012.

25. P. Mohd Arsad, N. Buniyamin, and J. L. Ab Manan, "Neural Network and Linear Regression methods for prediction of students' academic achievement," IEEE Glob. Eng. Educ. Conf. EDUCON, no. April, pp. 916–921, 2014.

26. G. Gray, C. McGuinness, and P. Owende, "An application of classification models to predict learner progression in tertiary education," in Souvenir of the 2014 IEEE International Advance Computing Conference, IACC 2014, 2014.

27. H. Hamsa, S. Indiradevi, and J. J. Kizhakkethottam, "Student Academic Performance Prediction Model Using Decision Tree and Fuzzy Genetic Algorithm," Procedia Technol., 2016.

28. Susan Bergin , Statistical and Machine Learning Models to Predict Programming Performance, Ph.D Thesis, Department of Computer Science, National University of Ireland, Maynooth, Ireland ,2006